



DIABETES DIAGNOSIS USING MACHINE LEARNING

Boshra Farajollahi^{1*}, Maysam Mehmannaavaz², Hafez Mehrjoo², Fateme Moghbeli³,
Mohammad Javad Sayadi¹

¹Department of Health Information Management, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran.

²Doornama Company, Data Science lab, Ilam, Iran.

³PhD of Medical Informatics, Assistant Professor, Department of HIT, Varastegan Institute for Medical Sciences, Mashhad, Iran.

Article Info

Article type:
Research

Article History:
Received: 2020-12-12
Accepted: 2021-03-04
Published: 2021-03-04

* Corresponding author:

Boshra Farajollahi
Department of Health Information
Management, School of Health
Management and Information
Sciences, Iran University of Medical
Sciences, Tehran, Iran.
Email:
boshrafarajollahi1373@gmail.com

ABSTRACT

Introduction:

Diabetes is a disease associated with high levels of glucose in the blood. Diabetes make many kinds of complications, which also leads to a high rate of repeated admission of patients with diabetes. The aim of this study is to diagnose Diabetes with machine learning techniques.

Material and Methods:

The datasets of the article contain several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age. The main objective of the machine learning models is to classify of the diabetes disease.

Results:

Six classifiers have been also adapted and compared their performance based on accuracy, F1-score, recall, precision and AUC. And Finally, Adaboost has the most accuracy 83%.

Conclusion:

In this paper a performance comparison of different classifier models for classifying diagnosis is done. The models considered for comparison are logistic regression, Decision Tree, support vector machine (SVM), xgboost, Random Forest and Adaboost. Finally, in the comparison flow, Adaboost, Logistic Regression, SVM and Random Forest, usually has had a high amount; and their amounts has little differences normally.

Keywords:

Diagnosis, Diabetes, Machine Learning.

How to cite this paper

Farajollahi B, Mehmannaavaz M, Mehrjoo H, Moghbeli F, Mahaki M. Diabetes Diagnosis Using Machine Learning. Front Health Inform. 2021; 10: 65. DOI: [10.30699/fhi.v10i1.273](https://doi.org/10.30699/fhi.v10i1.273)

INTRODUCTION

Diabetes is a chronic disease and commonly stated by health professionals or doctors as diabetes mellitus (DM), which describes a set of metabolic diseases in which the person has blood sugar, either insulin production inefficient, or because of the body cell do not return correctly to insulin, or by both reason [1]. This will increase concentration levels of glucose in the blood [2]. The majority of cases of diabetes can be broadly classified in two categories, type 1 and type 2, although some cases are difficult to classify [3]. Many complications occur if diabetes remain untreated [2]. Therefore, it is not only a disease but also a creator of different kinds of diseases like heart

attack, blindness, kidney diseases, etc. [1]. Diabetes has become one of the major causes of national disease and death in most countries [4]. According to the International Diabetes Federation report, this figure is expected to rise to more than 642 million in 2040, so early screening and diagnosis of diabetes patients have great significance in detecting and treating diabetes on time [4]. The analysis of diabetes data is a challenging issue because most of the medical data are nonlinear, abnormal, correlation structured, and complex in nature [5]. Applying machine learning methods in diabetes mellitus research is a key approach to utilizing large volumes of available diabetes-related data for extracting knowledge [2]. It also helps the people to accurately

diagnosis of diabetes [5]. The purpose of this study was to compare performance analysis of logistic regression (LR), decision tree (DT), support vector machine (SVM), xgboost, random forest (RF) and adaboost models to diabetes mellitus classification. In fact the purpose of using these algorithms is the comparison between different algorithms such as ensemble learning and linear classifiers.

MATERIAL AND METHODS

The dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. The datasets consist of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

Feature selection

Optimizing the performance of the classification model by feature selecting is an important part [6]. Feature selection, as a data preprocessing strategy, has been proven to be effective and efficient in preparing data (especially high-dimensional data) for various data mining and machine-learning problems. The objectives of feature selection include building simpler and more comprehensible models, improving data mining performance, and preparing clean, understandable data [7]. The Principal Components Analysis (PCA) was used for data reduction, which can not only greatly reduce the time of model learning while preserving the data implied information, but also eliminate data noise and data redundancy [8]. In this study, PCA-based feature selection and the best feature selection which has 7 features including: age, Skin Thickness, Glucose, Blood Pressure, Diabetes Pedigree Function, Insulin and Pregnancies, has been applied.

Machine learning models

The main objective of the machine learning models is to classify of the diabetes disease. The overview of the proposed machine learning models has been shown in Fig 1.

The training/test set paradigm of the entire machine learning models has been shown in Fig 2. The first step is to divide the dataset into two sets such as 80% for training set and 20% for testing set. The training and test sets are separated. In the second step, the most significant risk factors of diabetes disease have been selected based on PCA feature selection. We

have adopted six classifiers; logistic regression, decision tree, support vector machine (SVM), xgboost, random forest (RF) and adaboost. The next step is to estimate the training classifier coefficients, and then the test classifiers have been applied to classify the patients into two categories as diabetic vs. control. Finally, the performances of the classifiers are evaluated using five performance parameters, namely accuracy, F1-score, recall, precision and AUC.

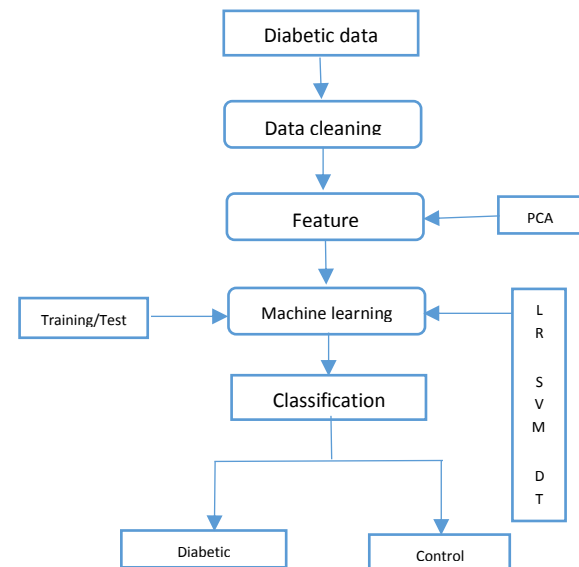


Fig 1: Overview of the proposed machine learning models

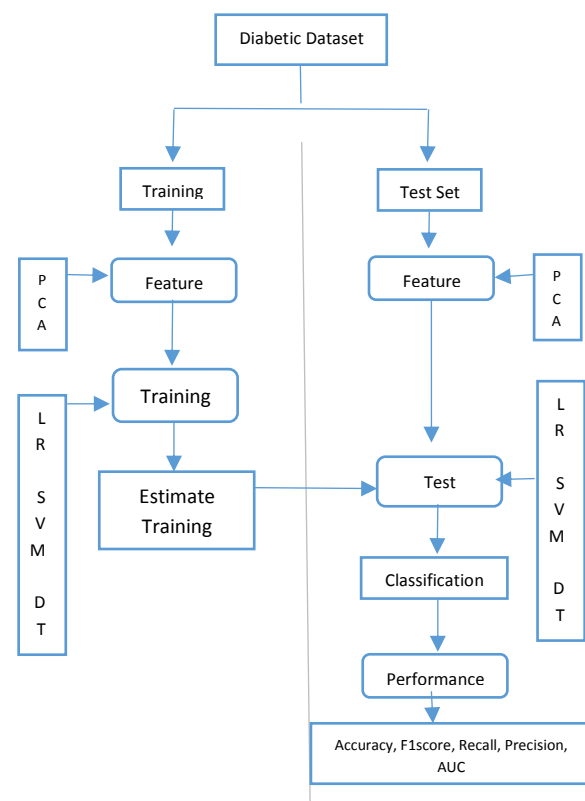


Fig 2: The training/test set paradigm of the machine learning models

RESULTS

In the section, we show the performance of machine learning classification techniques for diabetes classification. For this, we analyze various popular classification techniques that include the logistic regression, decision tree, support vector machine (SVM), xgboost, random forest and adaboost. The high risk factors have been selected based on PCA feature selection. Moreover, six classifiers have been also adapted and compared their performance based on accuracy, F1-score, recall, precision and AUC. The next section represents the related work.

Comparison of the Efficiency of Algorithms

Fig 3 demonstrates the comparison of the performance of 6 machine learning algorithms based on accuracy. Adaboost has the most accuracy and after that two algorithms, random forest and logistic regression, which has the same accuracy, have high precision. In the next rate, SVM with accuracy of 82.46 almost has a high accuracy. Decision tree and xgboost have accuracy below 80 percent. Finally xgboost has the least accuracy.

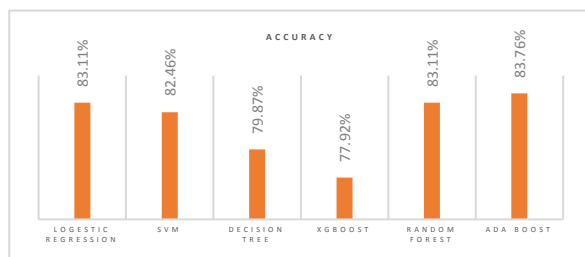


Fig 3: Comparison of the Performance of machine learning Algorithms based on Accuracy

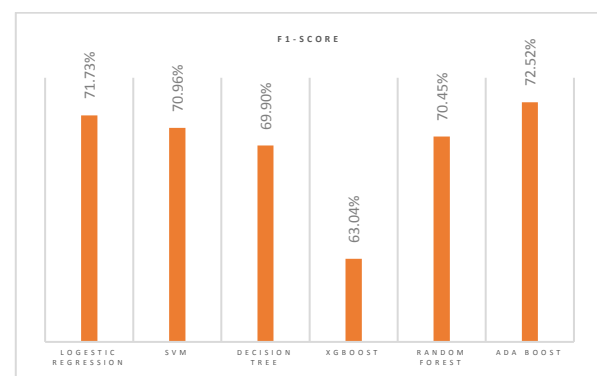


Fig 4: Comparison of the Performance of machine learning Algorithms based on F1-score

Fig 4 represents the comparison of the performance of 6 machine learning algorithms based on F1-score. Adaboost has the most F1-score. After that, logistic regression and SVM have high F1-score more than 70% respectively. F1-score in decision tree is less than 70% and xgboost has the least F1-score among algorithms.

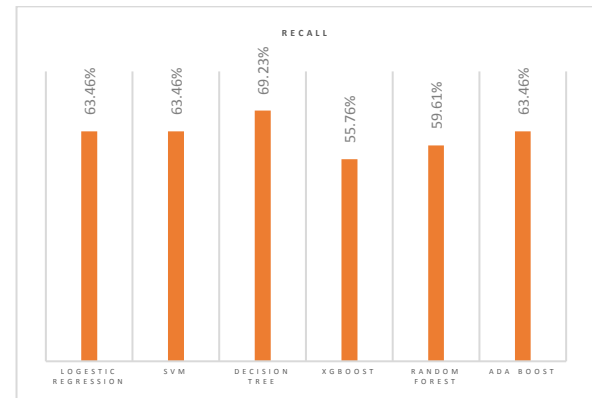


Fig 5: Comparison of the Performance of machine learning Algorithms based on Recall

Fig 5 indicates the comparison of the performance of 6 machine learning algorithms based on recall. Decision tree has the most recall among algorithms. Adaboost, SVM and logistic regression have equal recall. Four mentioned algorithms have a recall more than 60%. Xgboost and random forest have a recall less than 60%, and xgboost has the least recall among algorithms.

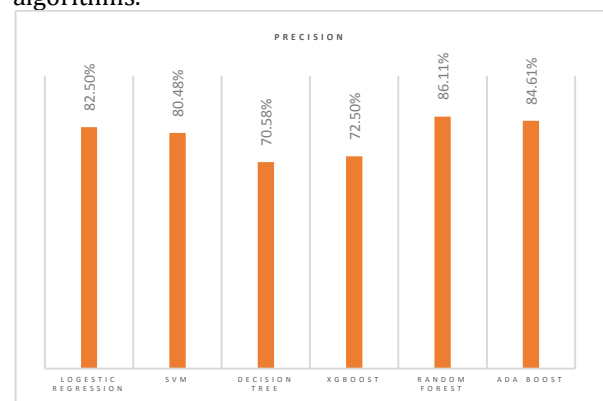


Fig 6: Comparison of the Performance of machine learning Algorithms based on Precision

Fig 6 points out the comparison of the performance of 6 machine learning algorithms based on precision. Random forest has the most precision. After that, adaboost, logistic regression and SVM have precision more than 80% respectively. Both decision tree and xgboost have precision less than 80%, and decision tree has the least precision among algorithms.

Receiver operating characteristics (ROC) is a graphical plot that is created by plotting sensitivity versus '1-specificity'. The area under the curve (AUC) which is computed from ROC curve is the indicator to evaluate the performance of the classifiers. The value of the AUC lies between '0' to '1' [5].

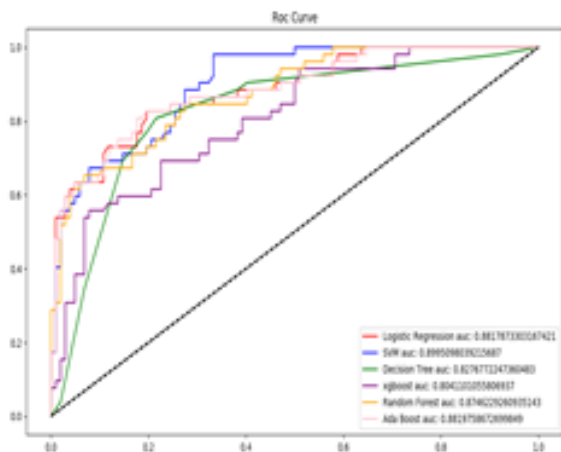


Fig 7: ROC curves of six classifiers

Fig 7 illustrates ROC curves of six classifiers. The amount of AUC of SVM is more than other algorithms, and after that, the amount of AUC in adaboost, logistic regression and Random forest is high respectively; which are close to each other, somehow. It can be said that almost the amount of AUC in decision tree and xgboost in related to the mentioned algorithms is diminished significantly, and it has the least AUC in xgboost.

The models considered for comparison are logistic regression, decision tree, support vector machine (SVM), xgboost, random forest and adaboost. The accuracy, F1-score, recall, precision and AUC are considered for the comparison. It can be stated that in the algorithms, in different comparison, xgboost has had the least amount and it has not just had the least precision; however, it has had a close precision to the least precision one. In the comparison flow, decision tree has had different positions. For instance, although it has had the least precision, it has had the most recall; and its amount in AUC, F1-score and accuracy is close to the least amounts. Altogether, it can be noted that in the comparison flow, adaboost, logistic regression, SVM and random forest, usually has had a high amounts; and their amounts has little differences normally (Fig 8).

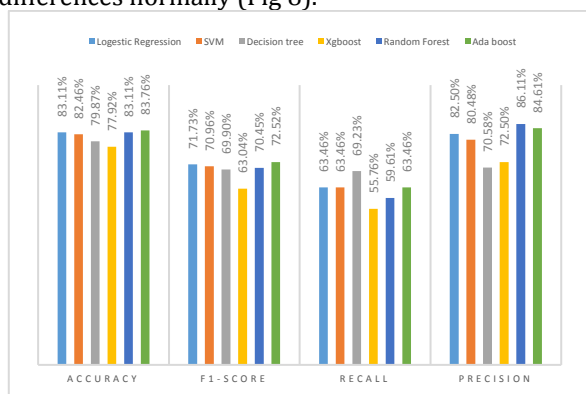


Fig 8: performance comparison results with accuracy, F1-score, recall and precision

DISCUSSION

These are the results of searches in scopus, google scholar and pubmed databases. The diagnosis of diabetes disease has been studied by Warke et al. In this project the primary aim is to analyses the diabetes dataset and use support vector machine, Naïve Bayes, logistic regression, and K-nearest neighbors algorithms; this analyses helps to predict and to develop a prediction engine. Developing a web application with following feature is the secondary aim. This project demonstrated a comparison of Naïve Bayes classifier with other linear classifiers, such as support vector machines, logistic regression, and K-nearest neighbors. The result of this project is that the chances of diabetes with more accuracy as compared to other classifiers, can be predicted by Naive Bayes machine learning classifier [1].

In essay [1] only the comparison of some kinds of linear algorithms has been done; while, in the current study, the comparison between some kinds of algorithms including linear classifier and ensemble learning is done.

Maniruzzaman et al. have published a project named "Classification and prediction of diabetes disease using machine learning paradigm". The main target of this study is to develop a machine learning (ML)-based system in order to predict diabetic patients. To identify the risk factors for diabetes disease based on p-value and odds ratio (OR), Logistic regression (LR) is used. Maniruzzaman et al. aimed to predict the diabetic patients; thus, they have adopted four classifiers like adaboost, Naïve Bayes (NB), decision tree (DT), and random forest (RF). Three kinds of partition protocols (K2, K5, and K10) have also chosen; which repeated these protocols into 20 trails. These classifiers' performances are evaluated by using accuracy (ACC) and by the area under the curve (AUC). The outcome was that the combination of LR and RF-based classifier performs are better; which will be very useful to predict diabetic patients [5].

"Diabetes diagnosis via XCS classifier system" is title of a project which has been published by Moshtaghi et al. In order to design an expert clinical system, this study aimed to use novel concepts of artificial intelligence. Diagnosing the diabetes disease automatically at the right time is the capability of this system. A learning system is the expert system which has been developed in this paper; it was as an improved version of extended classifier systems (XCS). The system in this research started to learn by application of a real dataset collected. The performance of that system was examined on some 268 other patients then. The results of that examination were compared with some conventional data mining methods. So as to predict accurately, this comparison indicates the preference of the proposed method with other techniques. The suggested

method has been applied (XCSR, AD Tree, SVM, C4.5, k star, Dempster-Saffr), and in test phase the results which were obtained from performing improved XCSR algorithm were compared with four other algorithms in the following table [9].

CONCLUSION

Diabetes mellitus is commonly known as diabetes. It is of group of metabolic orders which are characterized by the high blood sugar. Diagnosis of diabetes is an important real-world of medical problems. Detection of diabetes one way out before treatment. In this paper a performance comparison of different classifier models for classifying diagnosis is done.

AUTHOR'S CONTRIBUTION

The authors agree on this final form of the manuscript, and attested that all authors contributed in the final draft of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding the publication of this study.

FINANCIAL DISCLOSURE

No financial interests related to the material of this manuscript have been declared.

REFERENCES

1. Warke M, Kumar V, Tarale S, Galgat P, Chaudhari D. Diabetes diagnosis using machine learning algorithms. *International Research Journal of Engineering and Technology*. 2019; 6(3): 1470-6.
2. Kavakiotisab I, Tsave O, Salifoglou A, Maglaveras N, Vlahavasa I, Chouvarda I. Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*. 2017; 15: 104-16.
3. Benbelkacem S, Atmani B. Random forests for diabetes diagnosis. *International Conference on Computer and Information Sciences*. IEEE; 2019.
4. Sun YL, Zhang DL. Machine learning techniques for screening and diagnosis of diabetes: A survey. *Tehnički Vjesnik*. 2019; 26(3): 872-80.
5. Maniruzzaman M, Rahman MJ, Ahammed B, Abedin MM. Classification and prediction of diabetes disease using machine learning paradigm. *Health Inf Sci Syst*. 2020; 8(1): 7. PMID: 31949894 DOI: 10.1007/s13755-019-0095-z [PubMed]
6. Pujianto U, Setiawan AL, Rosyid HA, Salah AMM. Comparison of naïve Bayes algorithm and decision tree C4. 5 for hospital readmission diabetes patients using hba1c measurement. *Knowledge Engineering and Data Science*. 2019; 2(2): 58-71.
7. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, et al. Feature selection: A data perspective. *ACM Computing Surveys*. 2017; 50(6): 1-45.
8. Jia M, Tian F. Readmission prediction of diabetic based on convolutional neural networks. *International Conference on Computer and Communications*. IEEE; 2019.
9. Moshtaghi Yazdani N, Yazdani Seqeloo A. Diabetes diagnosis via XCS classifier system. *Iran Med Inform*. 2014; 3(1): 1-8.